ANÁLISIS CLUSTER

IDEA CONCEPTUAL BÁSICA:

- La heterogeneidad de una población constituye la materia prima del análisis cuantitativo.....
- ... sin embargo, en ocasiones, el individuo u objeto particular, aislado, resulta un "recipiente" de heterogeneidad demasiado pequeño,..... la unidad de observación es demasiado reducida con relación al objetivo del análisis....
- ... en estos casos, se trata entonces de agrupar a los sujetos originales en grupos, centrando el análisis en esos grupos, y no en cada uno de los individuos.....
- ... si existe una "taxonomía" ya diseñada que resulte útil, ajustada al objetivo de análisis, se recurre a ella,
- ... pero si no es así, deberemos crearla, generando una nueva "agrupación" que responda bien a las dimensiones de nuestro análisis.

DEFINICIÓN:

- **Se utiliza la información de una serie de variables para cada sujeto u objeto y, conforme a estas variables se mide la similitud entre ellos. Una vez medida la similitud se agrupan en: grupos homogéneos internamente y diferentes entre sí.**

- La "nueva dimensión" lograda con el cluster se aprovecha después para facilitar la aproximación "segmentada" de un determinado análisis.

CONVIENE TENER CLARO DESDE EL PRINCIPIO:

- Que la técnica no tiene vocación / propiedades inferenciales

- Que por tanto, los resultados logrados para una muestra sirven sólo para ese diseño (su valor atañe sólo a los objetivos del analista): elección de individuos, variables relevantes utilizadas, criterio similitud utilizado, nivel de agrupación final elegido.... **definen diferentes soluciones**.

- Que cluster y discriminante no tiene demasiado en común: el discriminante intenta explicar una estructura y el Cluster intenta determinarla.

ANÁLISIS CLUSTER

2 OBJETIVOS BÁSICOS:

- **Análisis "taxonómico"** con fines exploratorios o confirmatorios.
- **Cambio (simplificación) de la dimensión** de los datos (*lo descrito al inicio de este documento: agrupación de objetos individuales en nuevas estructuras de estudio (grupales))*

CLUSTER (Un ejemplo):

(Objetivo) Una empresa desea clasificar a sus consumidores en "tipos" según sus distintas percepciones de determinados atributos de la marca: CALIDAD GLOBAL, NIVEL SERVICIO, PRECIO, SERVICIO POSTVENTA Y VARIEDAD.

(Diseño) Para ello, se diseña una muestra con 100 compradores a los que cuestiona sobre su percepción, en una escala de intervalo, de las anteriores 5 características de los productos de la empresa.

(Resultado) La idea final consiste en diseñar distintas estrategias de promoción en función de sus diversos perfiles, si es que estos existen.

ANÁLISIS CLUSTER

ETAPAS DE UN ANÁLISIS CLUSTER

1.- SELECCIÓN DE LA MUESTRA DE DATOS

2.- SELECCIÓN y TRANSFORMACIÓN DE VARIABLES A UTILIZAR

3.- SELECCIÓN DE CONCEPTO DE DISTANCIA O SIMILITUD Y MEDICIÓN DE LAS MISMAS

4.- SELECCIÓN y APLICACIÓN DEL CRITERIO DE AGRUPACIÓN

5.- DETERMINACIÓN DE LA ESTRUCTURA CORRECTA (Elección del número de grupos)

ANÁLISIS CLUSTER

1.- SELECCIÓN DE LA MUESTRA

- **Adecuar al máximo la muestra al objeto de análisis**
- **Depuración de atípicos** (interesan elementos como miembros de grupos, no interesa la excesiva "individualidad")

2.- SELECCIÓN DE VARIABLES

CANTIDAD

- **No elegir variables indiscriminadamente:** RECORDAMOS: cada estructura se manifiesta en una serie de variables y cada grupo de variables revela, sólo, una determinada estructura.
- Resultado muy sensible a la inclusión de alguna variable irrelevante.
- La inclusión indiscriminada de variables aumenta la probabilidad de atípicos.

TRANSFORMACIÓN ?

- **Depende / Afecta a** muchas decisiones posteriores (medida de distancia / similitud empleada, por ejemplo)
- **Estandarización por variable:** aunque resulta útil para mediciones posteriores de distancia puede afectar al resultado del análisis y no se recomienda si las diferencias de medidas reflejan alguna cualidad natural de interés conceptual.
- **Estandarización por encuestado:** singular, pero en baterías de indicadores elimina patrones de respuesta en los sujetos, ofreciendo la importancia relativa de cada indicador.
- **Factorización:** puede resultar interesante factorizar previamente las variables y realizar el Cluster con factores en lugar de con variables.
- **El tipo de escala de medida** afectará a fases posteriores del procedimiento.

ANÁLISIS CLUSTER

3.- MEDIDAS DE SIMILITUD O DISTANCIA

TIPOS

A.- **CORRELACIÓN:** Se traslada el concepto tradicional de co-variación, de conexión entre variables, de "pautas" de transición (por ejemplo, el cálculo de un coeficiente de correlación) aplicándolo a las observaciones de los sujetos como si fuesen observaciones de variables.

B.- **Medidas de SIMILITUD / DISTANCIA:** Definen proximidad, no Covariación, y su elección (tipos) viene determinada por la escala de medida de las variables: binaria u ordinal o de intervalo/razón.

- **Medidas de distancia** para escalas ordinales, de intervalo o razón; amplia variedad,
- **Medidas de similitud para variables nominales binarias:** reciben el nombre de medidas de asociación

UNA ADVERTENCIA BÁSICA:

¡¡¡¡¡ El resultado final del Cluster depende radicalmente de de la medida de ASOCIACIÓN / SIMILITUD / DISTANCIA utilizada. Se recomienda, en cada contexto, observar empíricamente esas diferencias. !!!!!

ANÁLISIS CLUSTER

ALGUNAS MEDIDAS DE DISTANCIA

EUCLÍDEA (para "t" variables)

$$d_{ij} = \sqrt{\sum_{k=1}^{t} (X_{ik} - X_{jk})^2}$$

- **Problemas con las unidades de medida:** normalización previa de variables recomendable. *Ojo: en SPSS obtenemos por defecto su cuadrado*

MANHATTAN (o función de la distancia absoluta, o City-Block)

$$d_{ij} = \sum_{k=1}^{t} \left| X_{ik} - X_{jk} \right|$$

- **Problemas con la colinealidad.** *En SPSS esta medida aparece con el nombre de BLOCK*

FORMULACIÓN GENERAL DE POWER (s,r)

$$d_{ij} = \left(\sum_{k=1}^{t} \left(X_{ik} - X_{jk} \right)^s \right)^{1/r}$$

- *En SPSS aparece como Power. Su variante más clásica es la de Minkowski (s=r).*

ANÁLISIS CLUSTER

D ^{2} DE MAHALANOBIS

$$d_{ij} = \left(X_i - X_j\right)' \sum^{-1} \left(X_i - X_j\right)$$

Donde Xi y Xj son matrices fila (1 x p) de observaciones para cada sujeto y S es la matriz de varianzas - covarianzas de las variables consideradas.

- **Dos ventajas de la D ^{2}:**

1.- **Se consigue mitigar el problema de las unidades en la medida** en que cada variable entra en el cálculo de distnacia corregida por su variabilidad (*función del tamaño*)

2.- **Se elimina la información redundante.** La más correcta en caso de elevada multi - colinealidad.

ALGUNAS MEDIDAS DE ASOCIACIÓN

	VARIABLE				
INDIVIDUO	**A**	**B**	**C**	**D**	**E**
I	1	0	0	1	1
J	1	1	0	1	1
K	0	1	1	0	1

- Convenimos: (a) si los individuos I y J tienen la variable, (b) si el individuo I tiene la variable y J no, (c) el individuo J tiene la variable e I no, (d) los individuos I y J no tienen la variable y p = a + b + c + d

(*)* SEMEJANZA SIMPLE: (a+d) /p
JACARD: a / (a+b+c)
DICE: 2a/ (2a+b+c+)
(*)* RUSSELL Y KAO: a/p

ANÁLISIS CLUSTER

4.- ALGORITMO DE AGRUPACIÓN - DIVISIÓN PARA LA OBTENCIÓN DE CONGLOMERADOS

I. JERÁRQUICOS (ESTRUCTURA PROGRESIVA ARBOL)

I.A.- JERÁRQUICOS AGLOMERATIVOS

I.A.1.- Distancia mínima (single linkage)
I.A.2.- Distancia máxima (complete linkage)
I.A.3.- Distancia entre centros (centroid)
I.A.4.- Distancia mediana (median)
I.A.5.- Distancia promedio

- **simple (average linkage)**
- **entre grupos (between groups)**
- **intragrupos (within groups)**

I.A.6.- Método de Ward

I.B.- JERÁRQUICOS DIVISIVOS

I.B.1.- Por cálculo iterativo de centros
I.B.2.- Monothetic
I.B.3.- Polythetic

II. NO JERÁRQUICOS (K-MEDIAS):

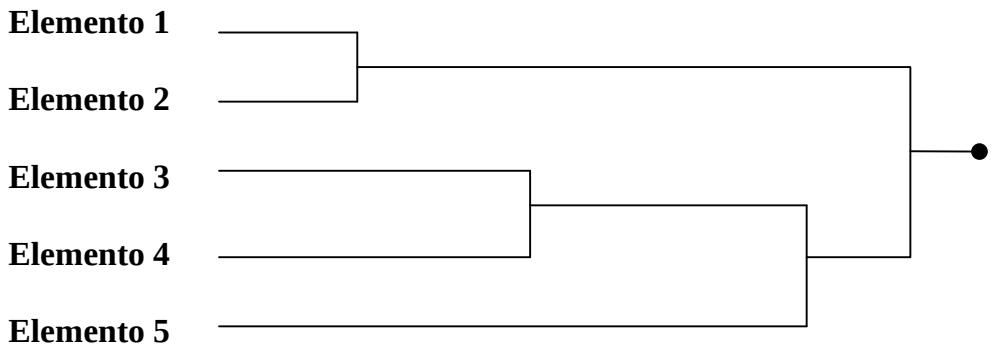
II.B.- UMBRAL SECUENCIAL
II.C.- UMBRAL PARALELO
III.D.- OPTIMIZACIÓN

ANÁLISIS CLUSTER

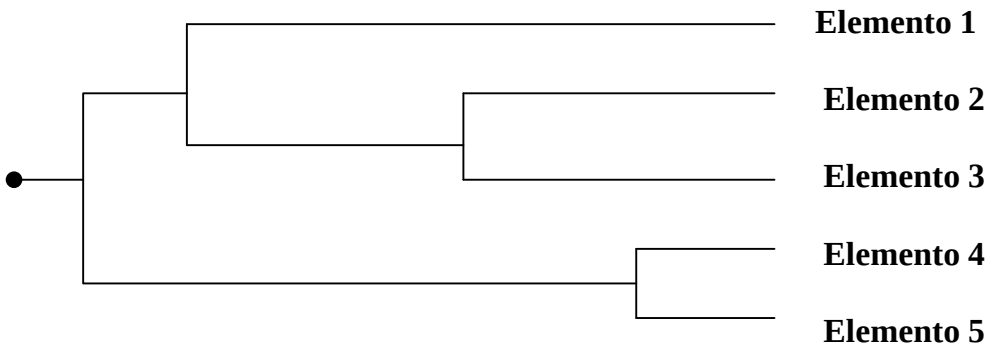
MÉTODOS JERÁRQUICOS

- **Definición:** la agrupación se realiza mediante proceso un con fases de agrupación o desagrupación sucesivas. El resultado final es una jerarquía de unión completa en la que cada grupo se une o separa en una determinada fase.

Método jerárquico **aglomerativo**:



Método jerárquico **divisivo**:



ANÁLISIS CLUSTER

DISTINTOS MÉTODOS AGLOMERATIVOS (Ejemplos)

- La selección de uno u otro método se basa en la forma en que la distancia se considera en el algoritmo de agrupación:

I.A.1.- Distancia mínima (single linkage)

Los grupos se unen considerando la menor de las distancias existentes entre los miembros más cercanos de distintos grupos.

(crea grupos más homogéneos pero permite cadenas de alineamientos entre sujetos muy lejanos)

I.A.2.- Distancia máxima (complete linkage)

Los grupos se unen considerando la menor de las distancias existentes entre los miembros más lejanos de distintos grupos.

(resuelve el anterior problema aunque los grupos son más heterogéneos)

I.A.6.- Método de Ward

IDEA BÁSICA: Se trata de ir agrupando de forma jerárquica elementos de modo que se minimice una determinada función objetivo.

FUNCIÓN A MINIMIZAR: Se perseguirá la minimización de la Variación Intra Grupal de la estructura formada.

(tiende a generar conglomerados demasiado pequeños y demasiado equilibrados en tamaño)

ANÁLISIS CLUSTER

$$SCI = \sum_{k=1}^{h} SCI_K$$

Partiendo de "h" grupos y "m" variables:

Para cada grupo

$$SCI_K = \sum_{i=1}^{m} \sum_{j=1}^{n_k} \left(X_{ijk} - \overline{X}_{ik} \right)^2$$

Suma cuadrática intra del grupo **"k"**

Suma de desviaciones en todas las variables (m) para todos los sujetos (n_j) dentro del grupo "k".

Valor de la variable "i" para cada sujeto "j" perteneciente al grupo "k" .

Media de la variable "i" en el grupo "k"

ANÁLISIS CLUSTER

DISTINTOS MÉTODOS NO JERÁRQUICOS (Ejemplos)

II.B.- UMBRAL SECUENCIAL

Se seleccionan una tras otra, "semillas" de conglomerado agrupando en torno a ellas todos los objetos que caen dentro de una determinada distancia. Cada objeto ya asignado no se considera para posteriores asignaciones.

II.B.- UMBRAL PARELELO

Similar al anterior pero se generan todas las semillas al mismo tiempo y los umbrales mínimas de aceptación en cada grupo.

III.D.- OPTIMIZACIÓN

Similares a los jerárquicos pero no se clasifican como tales porque en las etapas sucesivas se permite la reasignación de sujetos.

ANÁLISIS CLUSTER

5.- NÚMERO ÓPTIMO DE GRUPOS

- **No existen criterios objetivos y ampliamente válidos**

- **Hay una IDEA importante:** A medida que vamos formando grupos estos son menos homogéneos (las distancias para las que se forman los grupos iniciales son menores que las de los grupos finales)..... pero la estructura es más clara...

- **Por tanto, podemos fijar un OBJETIVO:** Identificar el punto de equilibrio entre la estructura incompleta y la estructura mezclada o confusa.....

- **No obstante, tenemos un problema.....:** Es difícil definir conceptualmente y más aún estadísticamente la situación de estructura correcta, no confusa, o la contraria de falta de estructura. (Estructura por asociación o diferenciación)....

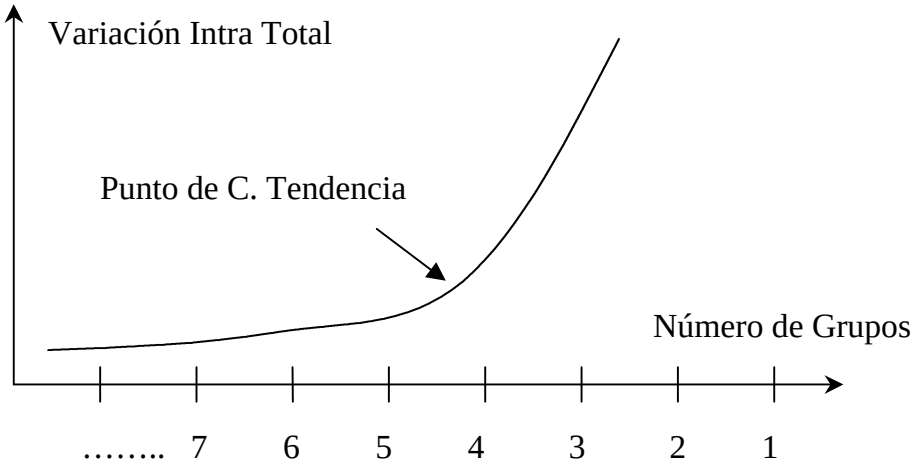
- **NOS APOYAREMOS, PARA DEFINIR LA ESTRUCTURA,** en la observación, tanto de las variables iniciales, como de la definición inicial de los sujetos y el significado de cada una de las etapas del proceso de agrupación.

- **Podemos, además, utilizar alguna herramienta técnica:** discriminante, caída brusca en la similitud o en la homogeneidad, ... dendograma, .

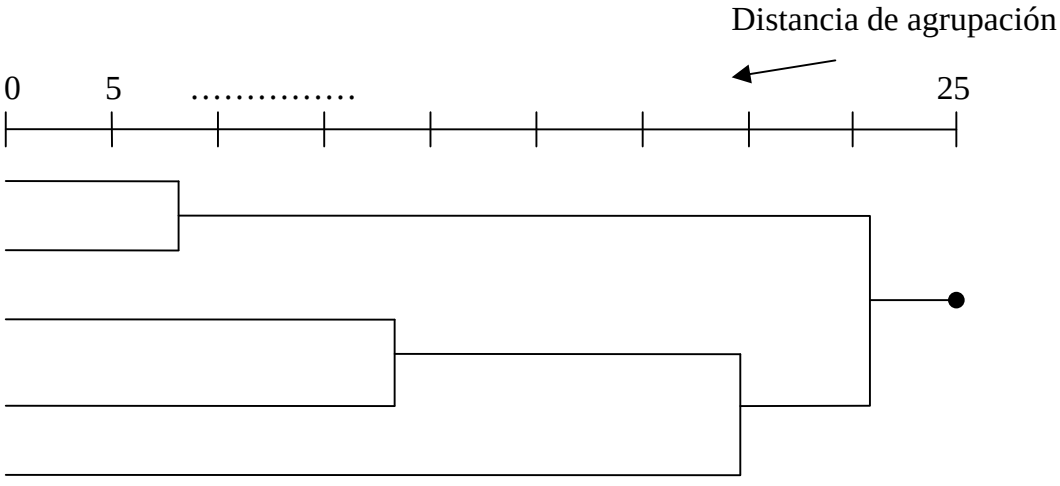
ANÁLISIS CLUSTER

TÉCNICAS DE AYUDA PARA DETERMINAR LA AGRUPACIÓN ÓPTIMA

- **Observación de la variación intragrupal**

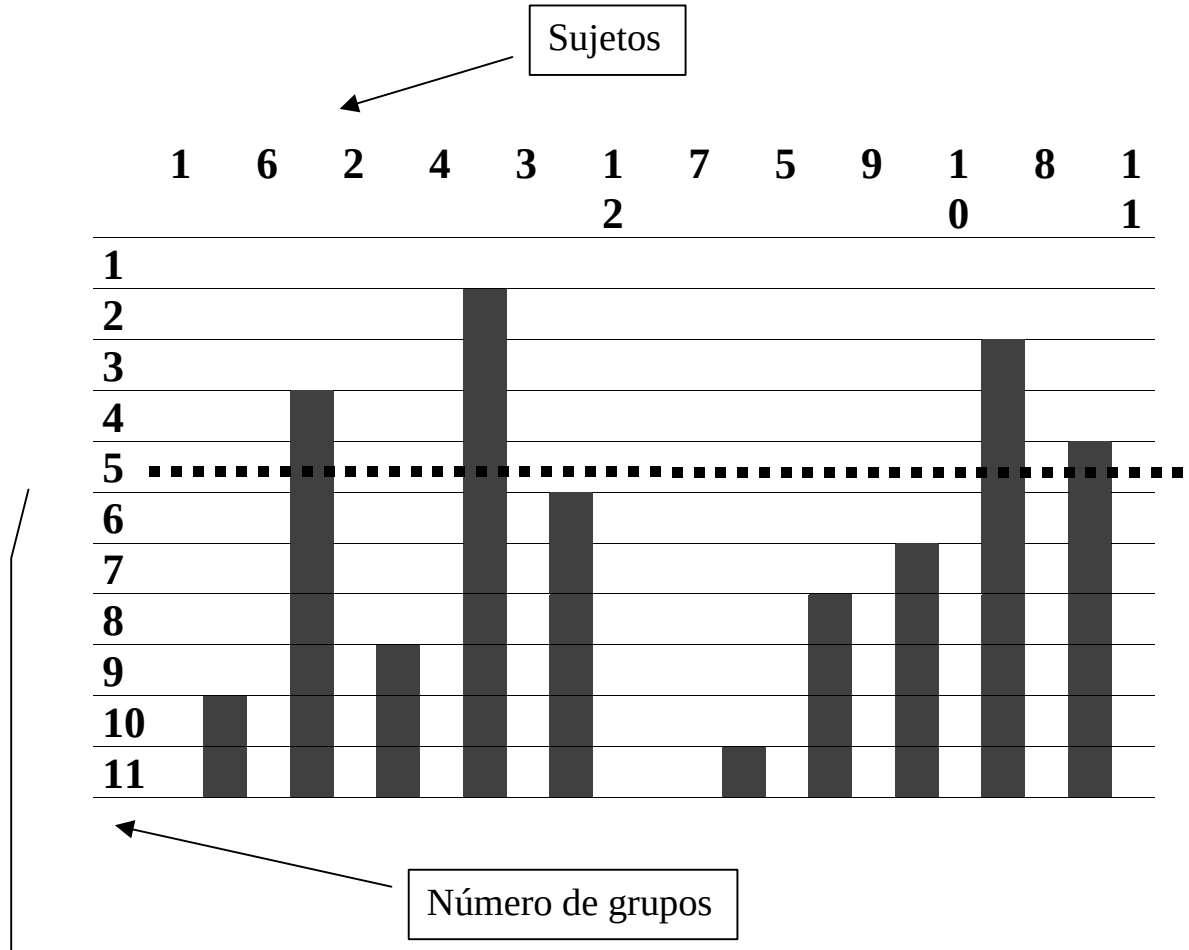


- **Dendograma**



ANÁLISIS CLUSTER

- **- Gráfico ICICLE**



Las barras negras delimitan la separación entre grupos. En el ejemplo, los cinco grupos son: (1,6) - (2,4) – (3,12,7,5,9,10) – (8) y (11)